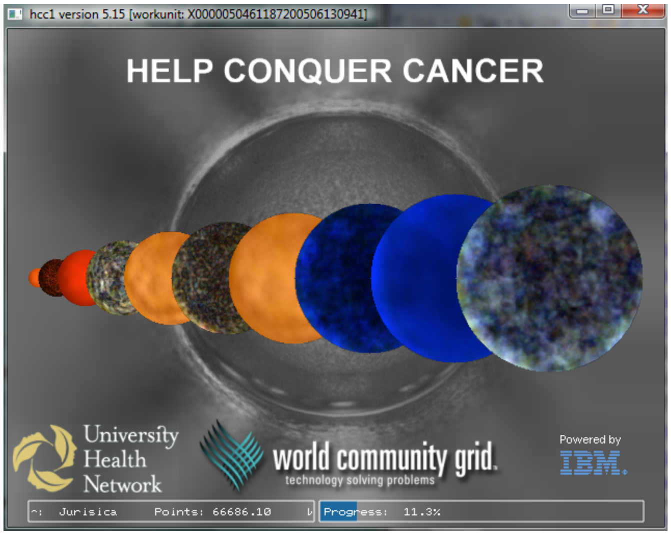
Update – January 2008


hcc1 version 5.15 [workunit: X0000050461187200506130941]

~: Jurisica Points: 66686.10

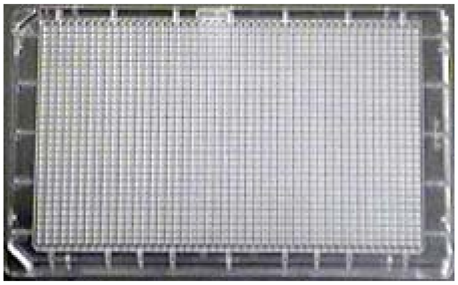
Progress: 11.3%

Dear WCG members,

first of all, we would like to thank you for supporting our project by donating your computational resources to tackle this complex problem.

Image Feature Extraction

Although we have over 84,000,000 images to process, we focused on a well-characterized set of 85,261 images first. All images come from over 9,400 protein screened at Hauptman Woodward Medical Research Institute (HWI) high-throughput screening facility, where robots are used to test each protein in 1,536 conditions. Each experiment is done in plates with 1,536 wells, as shown below:



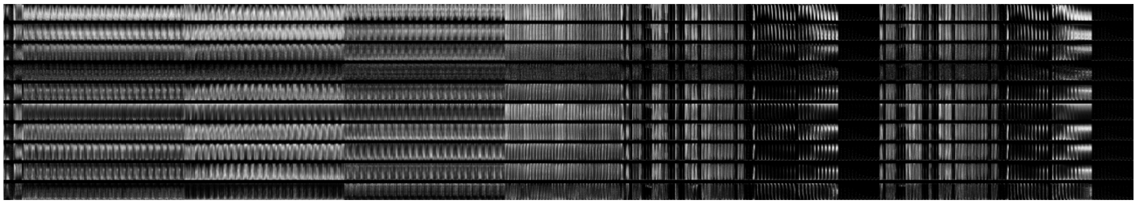
The high-priority set of images covers a wide range of potential outcomes, and has been extensively analyzed by human experts at HWI. WCG has already computed the image features for all of these images, resulting in an 85K by 12K data matrix.

This unique dataset is currently being analyzed to help us achieve several short term goals:

1. Extensively comparing results of over 12,000 image features extracted during image analysis across 10 outcome categories will enable us to determine strong association between feature values and outcome classes. This exploratory step was necessary, as computational complexity prohibited such comprehensive analysis on large number of images in the past.

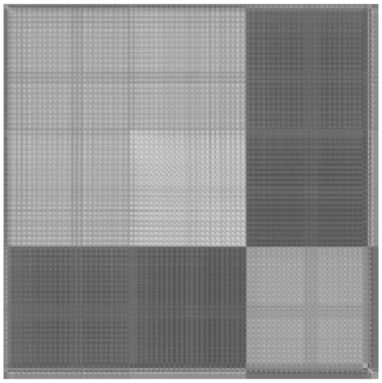
   A plot of feature-value distributions according to image class is shown below. Feature values were divided into 100 bins, arranged as much as possible so that images values were evenly distributed across bins. (Thus, bins = percentiles. The lowest 1% of feature values fall in bin 1, the next-lowest 1% into bin 2, and so forth.) Ignoring image class, each bin should be equally filled. The diagram shows distribution of feature values per image class (normalized per class). Heavy non-uniformity across image classes is apparent for many features. Each of 10 image classes is plotted as a horizontal band. Pixel-columns in this band correspond to different features. Brighter pixels indicate more populous bins. Image classes are ordered top-to-bottom as follows:

   a. Clear
   b. Phase separation
   c. Phase separation & precipitate
   d. Phase separation & skin
   e. Phase separation & crystal
   f. Precipitate
   g. Precipitate & skin
   h. Precipitate & crystal
   i. Crystal
   j. Garbage



2. Considering computational complexity, we also aim to eliminate features that are uninformative and features that have limited information value at a great computational cost. Analysis on a small set of images may lead to incorrect decision to choose one feature over the other. In the past, such decision was frequently guided by the goal to reduce CPU or memory requirement. With the help of WCG and the community, for the first time, we are now able to choose appropriate features based on extensive experimental evaluation.

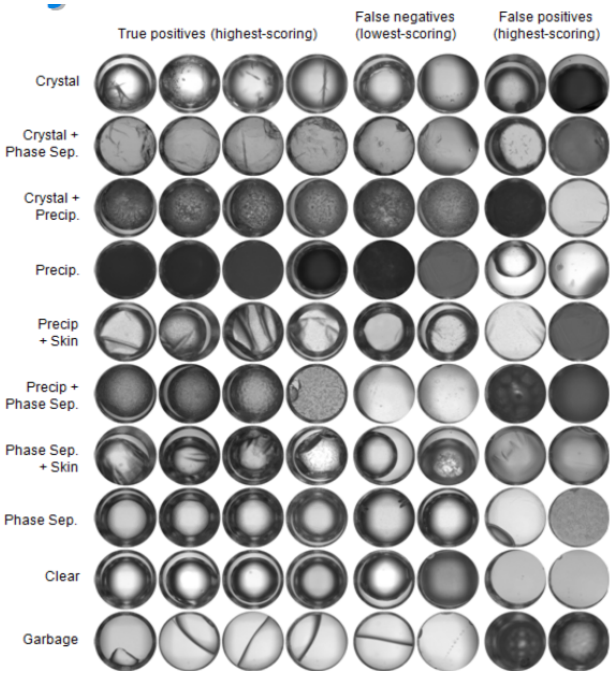
   A map showing correlation of computed features, ignoring image class is shown bellow. A complex relationship of strong correlation and anti-correlation exists.



3. Using the optimized feature set, we will update the automated crystallography image classifier, to enable follow up analysis, such as data mining to determine principles of crystal growth, and case-based reasoning to improve the crystallization plan optimization.

In the long run, we will use the updated crystallization image classification software to analyze the remaining images. This analysis will enable us to:

- Improve our understanding of the crystallization process. With information about the chemistry and outcomes of 14 million experiments involving 9,400 proteins (1,536 conditions per protein and 6 time points over period of 4 weeks) in a unified database we hope to be able to start learning some real chemistry about proteins.
- Discover proteins that have been screened and may have favorable crystallization conditions but were missed in the initial analysis. This requires that we update our initial classifier, results of which are shown below:



- Update the crystallization database, and compare any new protein screened to the results from over 9,400 proteins screened in the past to determine whether similar crystallization optimization may be applied to obtaining structure for the novel protein.

Target Selection

One can use several strategies to select targets for structural biology:

- **Cancer targets:** we use integrative cancer informatics to discover novel biomarkers for detecting disease early, prognostic markers, and markers that enable treatment optimization or measure response to treatment. We mostly focus on lung, ovarian, prostate, and head&neck cancers. Discovering the targets is only a first step. We need to extensively validate them; and since for many of them we do not have their structure, we need to determine it experimentally. Results of WCG computation will enable us to achieve this goal.
- **"Important" targets:** Protein structures are needed for many disease- or basic biology studies. To enable such research on a large scale, HWI facility screens proteins from around the word. To date, HWI helped over 800 labs. This very broad coverage is to ensure that all "important" proteins are considered.
- **"Extending the fold space coverage":** computationally selecting protein targets to maximize uniqueness of folds discovered.

Background

Protein Crystallography

Proteins are involved in every biochemical process that maintains life. Understanding protein structure will help us to understand functions of these important molecules. Correct protein function depends on their three-dimensional structure.

There are three main approaches to structure determination – *in silico* prediction, NMR (Nuclear Magnetic Resonance) and X-ray crystallography. Currently, the most powerful method for protein structure determination is single crystal X-ray diffraction, although new breakthroughs in NMR approaches are growing in their importance, and with more motifs and structures available the *in silico* methods are becoming more accurate.

A crystallography experiment begins with a well-formed crystal that ideally diffracts X-rays to high resolution. Despite the progress in experimental techniques, it remains challenging to crystallize novel proteins, and we still do not know most of the laws by which proteins adopt their three-dimensional structure. Thus, understanding these laws is one of the primary challenges in modern molecular biology.

Crystallization is a multi-parametric process with three classical steps: nucleation, growth and cessation of growth. Technical difficulties in protein crystallization are due to mainly two reasons:

- A large number of parameters affect the crystallization outcome, including purity of proteins, super-saturation, temperature, pH, time, ionic strength and purity of chemicals, volume and geometry of samples;

- We do not fully understand correlations between the variation of a parameter and the propensity for a given macromolecule to crystallize.

Conceptually, protein crystal growth can be divided into two phases: **search** and **optimization**. Search phase determines a subset of all possible crystallization conditions that yield promising crystallization outcome. These conditions are varied during the optimization phase to produce diffraction-quality crystals. Neither of the two phases is trivial to execute. If we consider only 15 possible conditions, each having 15 possible values, the result would be 4.3789e+017 possible experiments; impossible to test exhaustively. Even a broad search phase may not produce any promising conditions, and many of the promising leads may elude optimization strategies.

We can speed up the search phase and improve the optimization phase by applying high-throughput robotic screens with knowledge management. Eventually, discovering the principles of crystal growth should diminish protein crystallization as a bottleneck in modern structural biology.

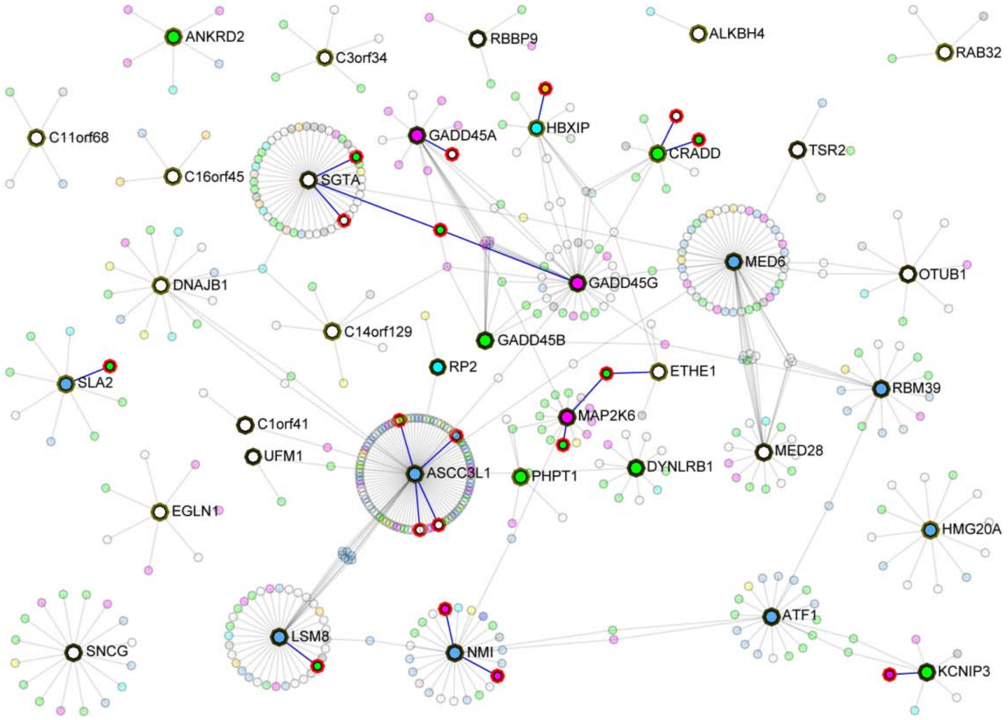
PDB - Protein DataBank

PDB is an international repository for protein structures (http://www.wwpdb.org/). It is important to improve our rate of protein structure determination, since so far PDB has only 37,404 structures available across multiple organisms. Considering that human has ~25,000 genes, most of which code for one or more proteins, we have a long way to go for human alone. It is the combination of better understanding the chemistry about proteins, and using robotics and information technology to streamline the process that will enable us to achieve this goal.

Target Selection – Integrative Cancer Informatics

Despite the introduction of many powerful chemotherapeutic agents over the past two decades, most cancers retain devastating mortality rates. To significantly impact cancer research, novel therapeutic approaches for targeting metastatic disease and diagnostic markers reflective of changes associated with disease onset that can detect early stage disease must be discovered. Better drugs must be rationally designed, and current drugs made more efficacious either by re-engineering or by information-based combination therapy. To tackle these complex biological problems and impact high-throughput biology requires integrative computational biology, i.e., considering multiple data types, developing and applying diverse algorithms for heterogeneous data analysis and visualization.

Computational biology will enable integrative analysis, visualization, interpretation, and modeling of these datasets. Protein interaction networks (http://ophid.utoronto.ca/i2d) will be used to define pathways, prioritize targets, reduce noise in high-dimensional screens, and plan network-based cancer targeting. Below we show one such network, using our network visualization software NAViGaTOR (http://ophid.utoronto.ca/navigator). To reduce graph complexity, we used alpha-blending, making nodes and edges translucent. Nodes represent proteins, their color represents biological function from GeneOntology, highlight color represents subset of structural biology targets currently under study in PSI (dark green), lung cancer markers (red). Edges represent protein interactions from I^{2}D, the highlighted edges represent direct interaction between lung cancer markers proteins with known or studied structure.


ANKRD2
C3orf34
RBBP9
ALKBH4
RAB32
C11orf68
C16orf45
SGTA
GADD45A
HBXIP
CRADD
TSR2
DNAJB1
GADD45G
MED6
OTUB1
C14orf129
GADD45B
RP2
SLA2
ETHE1
RBM39
MAP2K6
C1orf41
UFM1
ASCC3L1
PHPT1
DYNLRB1
MED28
EGLN1
HMG20A
SNCG
LSM8
NMI
ATF1
KCNIP3